`Prta`: A System to Support the Analysis of Propaganda Techniques in the News

Giovanni Da San Martino[1]   **Shaden Shaar**[1]   **Yifan Zhang**[1]
Seunghak Yu[2]   **Alberto Barrón-Cedeño**[3]   **Preslav Nakov**[1]
[1] Qatar Computing Research Institute, HBKU, Qatar
[2] MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
[3] Università di Bologna, Forlì, Italy
{gmartino,sshaar,yzhang,pnakov}@hbku.edu.qa
seunghak@csail.mit.edu
a.barron@unibo.it

Abstract

Recent events, such as the 2016 US Presidential Campaign, Brexit and the COVID-19 "infodemic", have brought into the spotlight the dangers of online disinformation. There has been a lot of research focusing on fact-checking and disinformation detection. However, little attention has been paid to the specific rhetorical and psychological techniques used to convey propaganda messages. Revealing the use of such techniques can help promote media literacy and critical thinking, and eventually contribute to limiting the impact of "fake news" and disinformation campaigns.

`Prta` (Propaganda Persuasion Techniques Analyzer) allows users to explore the articles crawled on a regular basis by highlighting the spans in which propaganda techniques occur and to compare them on the basis of their use of propaganda techniques. The system further reports statistics about the use of such techniques, overall and over time, or according to filtering criteria specified by the user based on time interval, keywords, and/or political orientation of the media. Moreover, it allows users to analyze any text or URL through a dedicated interface or via an API. The system is available online: https://www.tanbih.org/prta.

1 Introduction

Brexit and the 2016 US Presidential campaign (Muller, 2018), as well as major events such the COVID-19 outbreak (World Health Organization, 2020), were marked by disinformation campaigns at an unprecedented scale. This has brought the public attention to the problem, which became known under the name "fake news". Even though declared word of the year 2017 by Collins dictionary,[1] we find that term unhelpful, as it can easily mislead people, and even fact-checking organizations, to only focus on the veracity aspect.

At the EU level, a more precise term is preferred, *disinformation*, which refers to information that is both (*i*) false, and (*ii*) intents to harm. The often-ignored aspect (*ii*) is the main reasons why disinformation has become an important issue, namely because the news was *weaponized*.

Another aspect that has been largely ignored is the mechanism through which disinformation is being conveyed: using propaganda techniques. *Propaganda* can be defined as (*i*) trying to influence somebody's opinion, and (*ii*) doing so on purpose (Da San Martino et al., 2020). Note that this definition is orthogonal to that of disinformation: Propagandist news can be both true and false, and it can be both harmful and harmless (it could even be good). Here our focus is on the propaganda techniques: on their typology and use in the news.

Propaganda messages are conveyed via specific rhetorical and psychological techniques, ranging from leveraging on emotions —such as using *loaded language* (Weston, 2018, p. 6), *flag waving* (Hobbs and Mcgee, 2008), *appeal to authority* (Goodwin, 2011), slogans (Dan, 2015), and clichés (Hunter, 2015)— to using logical fallacies —such as *straw men* (Walton, 1996) (misrepresenting someone's opinion), *red herring* (Weston, 2018, p. 78),(Teninbaum, 2009) (presenting irrelevant data), *black-and-white fallacy* (Torok, 2015) (presenting two alternatives as the only possibilities), and *whataboutism* (Richter, 2017).

Here, we present `Prta` —the PRopaganda persuasion Techniques Analyzer. `Prta` makes online readers aware of propaganda by automatically detecting the text fragments in which propaganda techniques are being used as well as the type of propaganda technique in use. We believe that revealing the use of such techniques can help promote media literacy and critical thinking, and eventually contribute to limiting the impact of "fake news" and disinformation campaigns.

---

[1] https://www.bbc.com/news/uk-41838386

Technique ● Snippet
`loaded language` ● **Outrage** as Donald Trump suggests injecting disinfectant to kill virus.
`name calling, labeling` ● WHO: Coronavirus emergency is '**Public Enemy Number 1**'
`repetition` ● I still have a **dream**. It is a **dream** deeply rooted in the American **dream**. I have a **dream** that one day …
`exaggeration, minimization` ● Coronavirus '**risk to the American people remains very low**', Trump said.
`doubt` ● **Can the same be said for the Obama Administration**?
`appeal to fear/prejudice` ● **A dark, impenetrable and "irreversible" winter of persecution of the faithful by their own shepherds will fall**.
`flag-waving` ● Mueller attempts **to stop the will of We the People**!!! It's time to jail Mueller.
`causal oversimplification` ● **If France had not have declared war on Germany then World War II would have never happened.**
`slogans` ● **"BUILD THE WALL!"** Trump tweeted.
`appeal to authority` ● **Monsignor Jean-François Lantheaume, who served as first Counsellor of the Nunciature in Washington, confirmed that "Viganò said the truth. That's all."**
`black-and-white fallacy` ● Francis said these words: "**Everyone is guilty for the good he could have done and did not do … If we do not oppose evil, we tacitly feed it**.
`obfuscation, Intentional vagueness, Confusion` ● **Women and men are physically and emotionally different. The sexes are not "equal," then, and therefore the law should not pretend that we are!**
`thought-terminating cliches` ● **I do not really see any problems there.** Marx is the President.
`whataboutism` ● President Trump —**who himself avoided national military service** in the 1960's— keeps beating the war drums over North Korea.
`reductio ad hitlerum` ● "Vichy journalism," a term which now fits so much of the mainstream media. **It collaborates in the same way that the Vichy government in France collaborated with the Nazis.**
`red herring` ● **"You may claim that the death penalty is an ineffective deterrent against crime – but what about the victims of crime? How do you think surviving family members feel when they see the man who murdered their son kept in prison at their expense? Is it right that they should pay for their son's murderer to be fed and housed?"**
`bandwagon` ● He tweeted, "**EU no longer considers #Hamas a terrorist group. Time for US to do same.**"
`straw man` ● "Take it seriously, but with a large grain of salt." **Which is just Allen's more nuanced way of saying: "Don't believe it**."

Table 1: Our 18 propaganda techniques with example snippets. The propagandist span appears highlighted.

With `Prta`, users can explore the contents of articles about a number of topics, crawled from a variety of sources and updated on a regular basis, and to compare them on the basis of their use of propaganda techniques. The application reports overall statistics about the occurrence of such techniques, as well as their usage over time, or according to user-defined filtering criteria such as time span, keywords, and/or political orientation of the media. Furthermore, the application allows users to input and to analyze any text or URL of interest; this is also possible via an API, which allows other applications to be built on top of the system.

`Prta` relies on a supervised multi-granularity gated BERT-based model, which we train on a corpus of news articles annotated at the fragment level with 18 propaganda techniques, a total of 350K word tokens (Da San Martino et al., 2019).

Our work is in contrast to previous efforts, where propaganda has been tackled primarily at the article level (Rashkin et al., 2017; Barrón-Cedeño et al., 2019; Barrón-Cedeño et al., 2019). It is also different from work in the related field of computational argumentation, which deals with some specific logical fallacies related to propaganda, such as *ad hominem* fallacy (Habernal et al., 2018b).

Consider the game *Argotario*, which educates people to recognize and create fallacies such as *ad hominem*, *red herring* and *irrelevant authority*, which directly relate to propaganda (Habernal et al., 2017, 2018a). Unlike them, we have a richer inventory of techniques and we show them in the context of actual news.

The remainder of this paper is organized as follows. Section 2 introduces the machine learning model at the core of the `Prta` system. Section 3 sketches the full architecture of `Prta`, with focus on the process of collection and processing of the input articles. Section 4 describes the system interface and its functionality, and presents some examples. Section 5 draws conclusions and discusses possible directions for future work.

2 Data and Model

Data We train our model on a corpus of 350K tokens (Da San Martino et al., 2019; Yu et al., 2019), manually annotated by professional annotators with the instances of use of eighteen propaganda techniques. See Table 1 for a complete list and examples for each of these techniques.[2]

---

[2]Detailed list with definitions and examples is available at http://propaganda.qcri.org/annotations/definitions.html
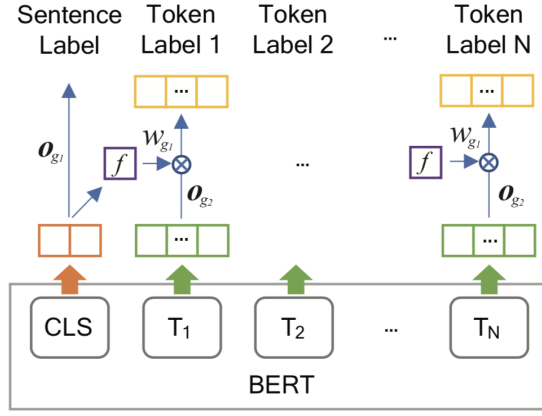
Figure 1: The architecture of our model.

Model Our model is based on multi-task learning with the following two tasks:

FLC *Fragment-level classification.* Given a sentence, identify all spans of use of propaganda techniques in it and the type of technique.

SLC *Sentence-level classification.* Given a sentence, predict whether it contains at least one propaganda technique.

Our model adds on top of BERT (Devlin et al., 2019) a set of layers that combine information from the fragment- and the sentence-level annotations to boost the performance of the FLC task on the basis of the SLC task. The network architecture is shown in Figure 1, and we refer to it as a multi-granularity network. It features 19 output units for each input token in the FLC task, standing for one of the 18 propaganda techniques or "no technique." A complementary output focuses on the SLC task, which is used to generate, through a trainable gate, a weight w that is multiplied by the input of the FLC task. The gate consists of a projection layer to one dimension and an activation function. The effect of this modeling is that if the sentence-level classifier is confident that the sentence does not contain propaganda, i.e., $w \sim 0$, then no propaganda technique would be predicted for any of the word tokens in the sentence.

The model we use in Prta outperforms BERT-based baselines on both at the sentence-level (F_1 of 60.71 vs. 57.74) and at the fragment-level (F_1 of 22.58 vs. 21.39). At the fragment-level, the model outperforms the best solution of a hackathon organized on this data.[3]

For the Prta system, we applied a softmax operator to turn its output into a bounded value in the range [0,1], which allows us to show a confidence for each prediction. Further details about the techniques, the model, the data, and the experiments can be found in (Da San Martino et al., 2019).[4]

3 System Architecture

Prta collects news articles from a number of news outlets, discards near-duplicates and finally identifies both specific propaganda techniques and sentences containing propaganda.

We crawl a growing list (now 250) of RSS feeds, Twitter accounts, and websites, and we extract the plain text from the crawled Web pages using the Newspaper3k library[5]. We then perform deduplication based on a combination of URL partial matching and content analysis using a hash function.

Finally, we use the model from Section 2 to identify sentences with propaganda and instances of use of specific propaganda techniques in the text and their types. We further organize the articles into topics; currently, the topics are defined using keyword matching, e.g., an article mentioning *COVID-19* or *Brexit* is assigned to a corresponding topic. By accumulating the techniques identified in multiple articles, Prta can show the volume of propaganda techniques used by each medium —as well as aggregated over all media for a specific topic— thus, allowing the user to do comparisons and analysis, as described in the next section.

4 Interface

Prta offers the following functionality.

For each crawled news article:

1. It flags all text spans in which a propaganda technique has been spotted.

2. It flags all sentences containing propaganda.

For a user-provided text or a URL:

3. It flags the same as in 1 and 2 above.

At the medium and at the topic level:

4. It displays aggregated statistics about the propaganda techniques used by all media on a specific topic, and also for individual media, or for media with specific political ideology.

Figure 2: Overall view for a topic.

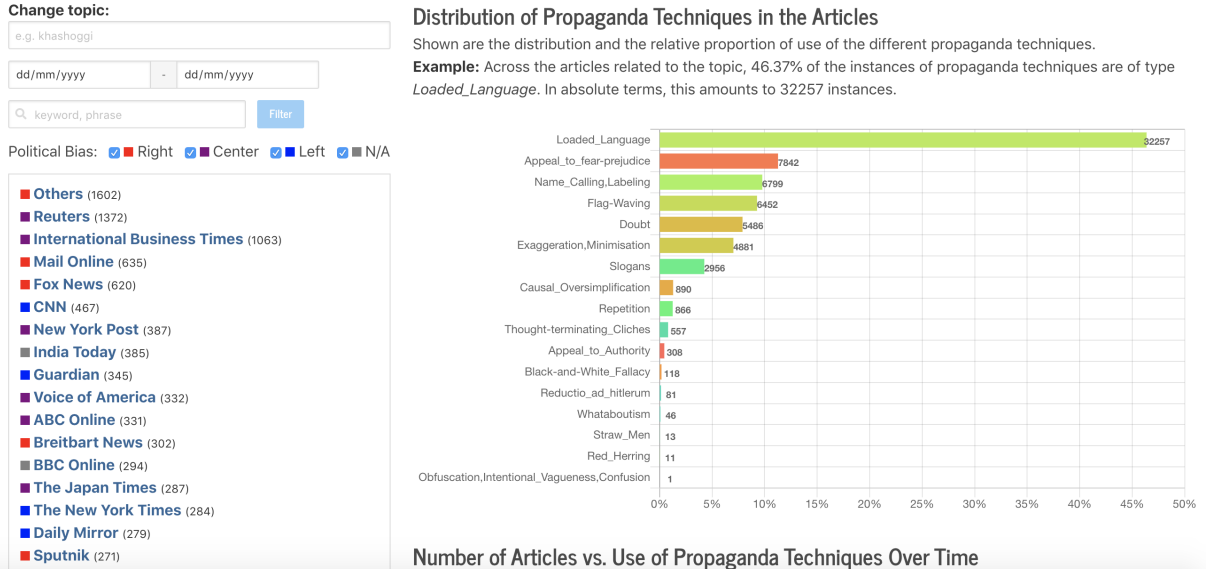## 4.1 Main Topic Page

This functionality is implemented in the three interfaces we expose to the user: the main topic page, the article page, and the custom article page, which we describe in Sections 4.1–4.3. Although points 1 and 2 above are run offline, they can also be invoked for a custom text using our API.[6]

Figure 2 shows a snapshot of the main page for a given topic: here, the *Coronavirus Outbreak in 2019-20*. We can see on the left panel, a list of the media covering the topic, sorted by number of articles. This allows the user to get a general idea about the degree of coverage of the topic by different media.

The right panel in Figure 2 shows statistics about the articles from the left panel. In particular, we can see the global distribution of the propaganda techniques in the articles, both in relative and in absolute terms. The right panel further shows a graph with the number of articles about the topic and the average number of propaganda techniques per article over time. Finally, it shows another graph with the relative proportion of propagandistic content per article; it is possible to click and to navigate from this graph to the target article. The latter two graphs are not shown in Figure 2, as they could not fit in this paper, but the reader is welcome to check them online.

The set of articles on the left panel can be filtered by time interval, by keyword, by political orientation of the media (left/center/right), as well as by any combination thereof.

Clicking on a medium on the left panel expands it, displaying its articles ranked on the basis of Eq. (1). Given the output of the multi-granularity network, we compute a simple score to assess the proportion of propaganda techniques in an article or in an individual media source. Let $F(x)$ be a set of fragment-level annotations in article x, where each annotation is a sequence of tokens. We compute the propaganda score for x as the ratio between the number of tokens covered by some propagandist fragment (regardless of the technique) and the total number of tokens in the article:

$$Q_a(x) = \frac{|\bigcup_{f \in F(x)} f|}{|x|}. \qquad (1)$$

Selecting a medium, or any other filtering criterion, further updates the graph on the center-right panel. For example, Figures 3a and 3b show the distribution of the techniques used by the BBC vs. Fox News when covering the topic of *Gun Control and Gun Rights*. We can see that both media use a lot of `loaded language`, which is the most common technique media use in general. However, the BBC also makes heavy use of `labeling` and `doubt`, whereas Fox News has a higher preference for `flag waving` and `slogans`.

---

[6]Link to the API available at https://www.tanbih.org/prta

(a) BBC on *Gun Control and Gun Rights*

(b) Fox News on *Gun Control and Gun Rights*
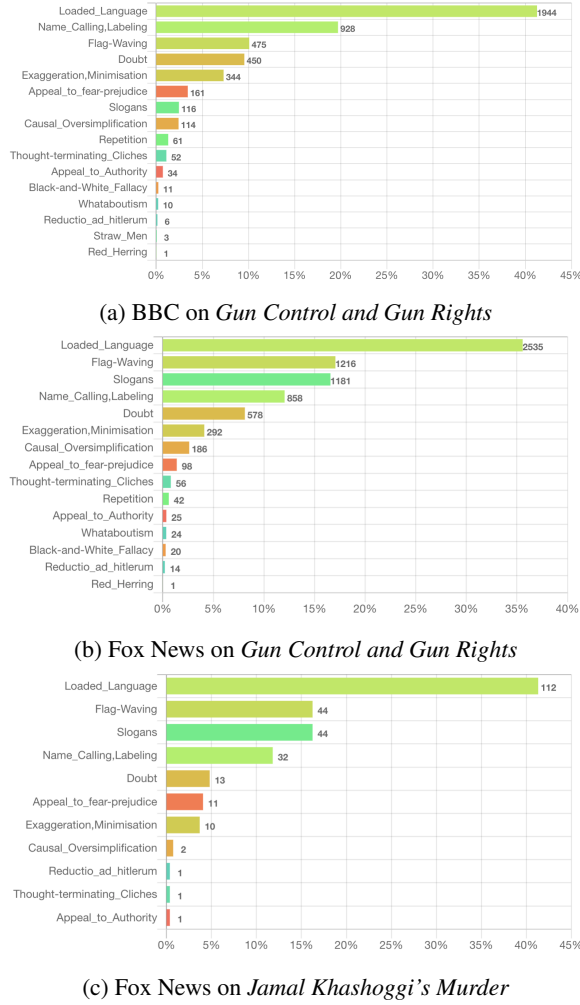

(c) Fox News on *Jamal Khashoggi's Murder*

Figure 3: Example of the distribution of the techniques as used by two media and on two different topics. Note that the scales are different.

Next, Figure 3c shows the propaganda techniques used by Fox News when covering the Khashoggi's Murder, which has a very similar technique distribution to the plot in Figure 3b.

This similarity between the distribution of propaganda techniques in Figures 3b and 3c might be a coincidence, or it could represent a consistent style, regardless of the topic. We leave the exploration of this and other hypotheses to the interested user, which is an easy exercise with the `Prta` system.

4.2 Article Page

When the user selects an article title on the left panel (Figure 2), its full content will appear on a middle panel with the propaganda fragments highlighted, as shown in Figure 4. Meanwhile, a right panel will appear, showing the color codes used for each of the techniques found in the article (the techniques that are not present are shown in gray).

Moreover, using the slider bar on top of the right panel, the user can set a confidence threshold, and then only those propaganda fragments in the article whose confidence is equal or higher than this set threshold would be highlighted. When the user hovers the mouse over a propagandist span, a short description of the technique would pop up. If the user wishes to find more information about the propaganda techniques, she can simply click on the corresponding question mark in the right panel.

4.3 Custom Article Analysis

Our interface allows the user to submit her own text for analysis. This allows her to find the techniques used in articles published by media that we do not currently cover or to analyze other kinds of texts. Texts can be submitted by copy-pasting in the text box on top, or, alternatively, by using a URL. In the latter case, the text box will be automatically filled with the content extracted from the URL using the Newspaper3k library (see Section 3), but the user can still edit the content before submitting the text for analysis. The maximum allowed length is the one enforced by the browser. Yet, we recommend to keep texts shorter than 4k in order to avoid blocking the server with too large requests.

Figure 5 shows the analysis for an excerpt of Winston Churchill's speech on May 10, 1940. All the techniques found in this speech are highlighted in the same way as described in Section 4.2. Notice that, in this case, we have set the confidence threshold to 0.4 and some of the techniques are consequently not highlighted. We can see that the system has identified heavy use of propaganda techniques. In particular, we can observe the use of *Flag Waving* and *Appeal to Fear*, which is understandable as the purpose of this speech was to prepare the British population for war.

5 Conclusion and Future Work

We have presented the `Prta` system for detecting and highlighting the use of propaganda techniques in online news. The system further shows aggregated statistics about the use of such techniques in articles filtered according to several criteria, including date ranges, media sources, bias of the sources, and keyword searches. The system also allows users to analyze their own text or the contents of a URL of interest.

We have made publicly available our data and models, as well as an API to the live system.

BREXIT

Figure 4: Selecting an article from the left panel, loads it and highlights its propaganda techniques.

SUBMIT YOUR TEXT

Figure 5: Analysis of a custom text, an excerpt from a speech by W. Churchill at the beginning of World War II. The confidence threshold is set to 0.4, and thus fragments for which the confidence is lower are not highlighted.

We hope that the `Prta` system would help raise awareness about the use of propaganda techniques in the news, thus promoting media literacy and critical thinking, which are arguably the best long-term answer to "fake news" and disinformation.

In future work, we plan to add more media sources, especially from non-English media and regions. We further want to extend the tool to support other propaganda techniques such as *cherry-picking* and *omission*, among others, which would require analysis beyond the text of a single article.

Acknowledgments

---

[7]http://propaganda.qcri.org
[8]http://tanbih.qcri.org

References

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, AAAI '19, pages 9847–9848, Honolulu, HI, USA.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, IJCAI-PRICAI '20, Yokohama, Japan.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '2019, Hong Kong, China.

Lavinia Dan. 2015. Techniques for the Translation of Advertising Slogans. In *Proceedings of the International Conference Literature, Discourse and Multicultural Dialogue*, LDMD '15, pages 13–23, Mures, Romania.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, MN, USA.

Jean Goodwin. 2011. Accounting for the force of the appeal to authority. In *Proceedings of the 9th International Conference of the Ontario Society for the Study of Argumentation*, OSSA '11, pages 1–9, Ontario, Canada.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 7–12, Copenhagen, Denmark.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018a. Adapting serious game for fallacious argumentation to German: pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC '18, Miyazaki, Japan.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 386–396, New Orleans, LA, USA.

Renee Hobbs and Sandra Mcgee. 2008. Teaching about propaganda: An examination of the historical roots of media literacy. *Journal of Media Literacy Education*, 6(62):56–67.

John Hunter. 2015. Brainwashing in a large group awareness training? The classical conditioning hypothesis of brainwashing. Master's thesis, University of Kwazulu-Natal, Pietermaritzburg, South Africa.

Robert Muller. 2018. Indictment of Internet Research Agency. https://commons.wikimedia.org/wiki/File:Internet_research_agency_indictment.pdf.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2931–2937, Copenhagen, Denmark.

Monika L Richter. 2017. The Kremlin's platform for 'useful idiots' in the West: An overview of RT's editorial strategy and evidence of impact. Technical report, Kremlin Watch.

Gabriel H Teninbaum. 2009. Reductio ad Hitlerum: Trumping the judicial Nazi card. *Michigan State Law Review*, page 541.

Robyn Torok. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. In *Proceedings of the Australian Security and Intelligence Conference*, pages 58–65, Perth, Australia.

Douglas Walton. 1996. *The straw man fallacy*. Royal Netherlands Academy of Arts and Sciences.

Anthony Weston. 2018. *A rulebook for arguments*. Hackett Publishing.

World Health Organization. 2020. Novel coronavirus (2019-ncov): situation report, 13. https://apps.who.int/iris/handle/10665/330778. Accessed: 2020-04-01.

Seunghak Yu, Giovanni Da San Martino, and Preslav Nakov. 2019. Experiments in detecting persuasion techniques in the news. In *Proceedings of the NeurIPS 2019 Joint Workshop on AI for Social Good*, Vancouver, Canada.